NAVAL MEDICAL RESEARCH UNIT– DAYTON

The Use of Commericial Flight Simulation Software as a Psychometrically Sound, Ecologically Valid Measure of Fatigued Performance

J. F. Chandler, D. S. Horning, R. D. Arnold,
J. B. Phillips, D. S. Horak, & D. L. Taylor

NAMRU-D REPORT NUMBER 11-34

NAVY MEDICINE
World Class Care…Anytime, Anywhere

Reviewed and Approved
12 August 2011

RG

Rita G. Simmons, CDR, MSC, USN
Commanding Officer, Acting

NAVAL MEDICAL RESEARCH UNIT
NAMRU DAYTON

The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

This work was funded by work unit number 70803.

The study protocol was approved by the Naval Aerospace Medical Research Laboratory Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects.

EXECUTIVE SUMMARY

Background

Fatigue is a deadly problem for U.S. Naval Aviation (Naval Safety Center, 2006), and receives a correspondingly large amount of research attention in military RDT&E. Though the *basic* consequences of fatigue are well known, significant measurement challenges remain in the applied laboratory, where an optimal combination of scientific rigor and operational relevance can be elusive. Completing a flight simulation (FS) while fatigued is used to maximize ecological validity for the flight environment (Caldwell et al., 2003; Russo et al., 2005; Van Dongen, Caldwell, & Caldwell, 2006), but this approach has psychometric shortcomings. The Psychomotor Vigilance Task (PVT) has significant psychometric strength as the gold standard instrument for assessing the cognitive effects of fatigue, but lacks ecological validity for flight. There is a clear need for a fatigue assessment tool combining the operational utility of a flight simulator with the control of the PVT.

Purpose

The purpose of this report is twofold: 1) to describe the development and execution of a flight simulation tool for quantifying vigilance during a fatigue study, and 2) to describe the effort to balance the control and diagnosticity of the PVT with the ecological validity of flight simulation in an inexpensive, off-the-shelf, open-source format.

Method

Fifteen active duty military personnel from the Naval Aviation Preflight Indoctrination (API) program at NAS Pensacola volunteered for the study. Subjects completed a battery of neurocognitive and physiological assessments over the course of 25 hours of continual wakefulness. As part of that battery, subjects completed eight trials of a simple flight profile using X-Plane 9 (Laminar Research, Columbia, SC), an inexpensive, off-the-shelf flight simulator. Subjects were instructed to fly “straight and level” at a specified altitude, airspeed and heading (i.e., 2000 ft, 140 knots, due North) for 15 minutes each session. To monitor second-by-second performance on the task, a central data capture server was connected to the four PC-based flight simulation stations. Deviations from the specified flight parameters were monitored by the data capture server. Lapse times were calculated for each parameter as the number of seconds during a simulator trial that subjects deviated from the flight goal by greater than one intraindividual standard deviation (determined at baseline). Flight Simulator Total Lapse Time was the sum of lapse times for each parameter. A series of Visual Basic (VB) programs were then written in Microsoft Excel to calculate descriptive statistics based on those lapses. Results were compared to concurrent subject performance on the PVT.

Results

Significant inter-trial correlations for the Flight Simulator Performance Task (FSPT) scores were moderately strong (r = .53 - .81, p < .05), providing initial evidence for test-retest reliability. However, there were some notable non-significant correlations which are discussed in terms of significant, trait-like individual differences in the data. These differences are quantified by demonstrating a significant inter-class correlation (ICC) value of .48, and a Cronbach’s Alpha of .89. Significant performance decrements on the FSPT were observed across time (p < .05), offering preliminary evidence of its construct validity as a fatigue assessment. Further, performance on the FSPT was able to successfully predict PVT lapses using Hierarchical Linear Modeling (HLM) (p < .05), indicative of convergent validity with a well-established fatigue assessment tool.

Discussion

Preliminary evidence suggests that performance on a simple flight simulation task can be used as a reliable, ecologically valid measure of fatigue in Student Naval Aviators. Future work should focus on replication and extension of the FSPT to further establish the measure’s psychometric properties.

INTRODUCTION

Fatigue's deadly influence on military and civilian aviation operations has been well established. Fatigue is cited as the primary or contributing causal factor in more military flight mishaps than any other causal factor, with predictable results; a significant cost in lost lives, money, and operational and training time (Naval Safety Center, 2006). Accordingly, fatigue in the cockpit has received a great deal of research attention over the past 50 years (see Caldwell et al., 2009). Yet significant measurement challenges remain in the applied laboratory, where achieving an optimal combination of scientific rigor and operational fidelity can be elusive.

The Psychometric Vigilance Task (PVT) has long been considered the gold standard instrument for assessing the cognitive effects of fatigue (Balkin et al., 2004). The task has been used extensively in applied fatigue research settings, including flight performance (see Dorrian, Rogers, & Dinges, 2005, for a review), but its simplicity limits researchers' ability to make strong inferences concerning its applicability to the flight environment. While the PVT captures vigilance in its basic form and offers the psychometric reliability missing in flight simulation (Dinges & Powell, 1985), the major drawback of the PVT is the lack of ecological validity for flight. For example, a performance lapse on the PVT (quantified as non-response to a randomly presented visual stimulus for more than 500 ms) may not translate well to a performance lapse in the cockpit. Sources of stimulation in the cockpit are numerous compared to the single focal point of the PVT, and may result in differential patterns of wake-state instability and cognitive resource allocation.

Conducting flight-based fatigue research in the cockpit under actual flight conditions would allow maximal ecological validity; but this concept poses countless logistical, measurement, and safety of flight related challenges. Flight simulation (FS) has the advantage of maximizing ecological validity vis-à-vis the flight environment (Caldwell et al., 2003; Russo et al., 2005; Van Dongen, Caldwell, & Caldwell, 2006), but this approach also presents psychometric challenges resulting from the use of multiple, disparate scoring approaches. The most commonly measured variables in FS fatigue research are adherence to a specific altitude, heading, and airspeed. These three core parameters are best suited to quantify the effects of fatigue on flight performance as they require attentional control with minimal stimulation; yet, methods to score *adherence to these parameters* vary widely. Several different scales and data manipulation techniques have been employed in an attempt to quantify fatigue-related performance impacts. These include, in ascending order of psychometric strength; measurement by flight instructors using a visual analog scale (Leino et al., 2007; Caldwell, Caldwell, Brown, & Smith, 2004) or interval scale (Leino et al., 2007), *Z*-scores (Adamson et al., 2010), magnitude of deviation from desired value (Caldwell, Smythe, Leduc, & Caldwell, 2000), and deviation calculated by root mean square error (Dalecki Bock, & Guardiera, 2010; Van Dongen, Caldwell, & Caldwell, 2006; Caldwell, Caldwell, Brown, & Smith, 2004; Caldwell et al., 2003).

Visual analog and interval scales rely on the interpretation of observers, which can result in an instrumentation threat to internal validity. The use of Z-scores (though more psychometrically sound than visual analog or interval scales) is based on group averaging of the performance data. This is problematic due to mounting evidence that large individual differences in fatigue resistance exist (Van Dongen, Baynard, Maislin, & Dinges, 2004). Magnitude of deviation from a desired value has the advantage of an easily interpretable composite outcome (i.e., an "accuracy" rating from 0 to 100). However, the major disadvantage of this approach is that it results in restriction of naturally occurring variance. Variance that is eliminated during the averaging process may reflect important differences between individuals. Root mean square (RMS) errors may be calculated to control for variability in simulator performance data and allow direct comparison of performance among various individual flight maneuvers or parameters. Though possessing certain advantages, this approach lacks capabilities inherent in measures such as the PVT whose known and stable psychometric properties facilitate comparisons between individuals, or with other measures of performance.

There is a clear need for a fatigue assessment tool combining the excellent measurement characteristics and control of the PVT with the operational utility of a flight simulator. Ideally, constructing and scoring a simulator-based fatigue performance task similar to the PVT would maximize the practical strengths of existing approaches while minimizing their psychometric weaknesses. The current approach combines measurement of magnitude of deviation from a desired value (failure to maintain specified flight parameters or lapses) with an individualized performance scoring approach theoretically similar to RMS error (departure from one's own baseline performance) while using a readily interpretable operational metric (lapses quantified in seconds). The purpose of this report is twofold: 1) to describe the development and execution of a flight simulation tool for quantifying the effects of fatigue on flight performance, and 2) to describe the effort to balance the control and diagnosticity of a PVT-like measure with the ecological validity of flight simulation in an inexpensive, off-the-shelf, open-source format.

METHOD

Subjects

Fifteen active duty military personnel from the Naval Aviation Preflight Indoctrination (API) program volunteered as test subjects as part of a larger fatigue study. The study protocol was approved by the Naval Aerospace Medical Research Laboratory Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects. Descriptive statistics for the subjects are presented in Table 1.

No specific groups were excluded. However, certain factors identified with a confidential medical history form served to exclude individual participants, due to their potential confounding effects. These included excessive alcohol use within the previous 48 hours (>3 drinks), greater than 400 mg of routine daily caffeine consumption, habitual use of tobacco products within the previous six months, and history of significant medical, neurological, psychiatric, or sleep-related problems (Killgore et al., 2009).

Table 1. Descriptive Statistics

	Age (years)		Height (in)		Weight (lbs)	
	Mean	SD	Mean	SD	Mean	SD
Male (n = 13)	24.7	2.1	71.2	3.3	186.6	20.0
Female (n = 2)	21.5	0.7	66.5	3.5	142.5	17.7
Total	24.3	2.3	70.5	3.6	180.7	24.6

Ethnicity	White	Black	Asian American	Hispanic/Latino(a)	Other
	11	2	0	2	0

Fatigue Assessments

Psychomotor Vigilance Task. The PVT-192 (Ambulatory Monitoring Inc., Ardsley, New York) is a brief vigilance and attention task, considered to be the gold standard instrument for assessing the effects of fatigue (Balkin et al., 2004). During each 10-minute trial, subjects are required to attend closely to a stimulus window and respond to the appearance of numbers by pressing a response button. Subjects are instructed to respond as quickly as possible. While the PVT provides numerous performance metrics, the score of interest in the current report is lapses (responses latencies of greater than 500 ms).

Flight Simulator Performance Task (FSPT). Simulated flight performance was measured using the X-Plane 9 (Laminar Research, Columbia, SC) flight simulator. Because fatigue impairs basic attentional processes, tasks which are subject to more reliable measurement were the focus of simulated flight performance. Specifically, subjects were given a simple flight profile, with instructions to fly "straight and level" at a specified altitude, airspeed and heading (i.e., 2000 ft, 140 knots, due North). Deviations from these specified flight parameters were assessed.

Design

The experiment employed a repeated measures design to investigate the effects of sleep deprivation on task performance over time. The experiment consisted of two phases, (1) the Practice Phase and (2) the Experimental/Sleep Deprivation Phase.

Procedures

Practice Phase. Up to four (4) volunteers were recruited during each week of the study. After receipt of participants' informed consent, the Practice Phase of the experiment began. This phase was executed Monday and Tuesday morning and required approximately 90 minutes of participation each day. Practice Phase data was used for each of the measures to establish performance asymptote and to mitigate practice effects during the Experimental/Sleep Deprivation phase. Each day participants completed 2 trials of the PVT and one 15-minute trial of the FSPT. (Participants completed other measures not relevant to this analysis as part of the larger study. Full details are reported in Chandler et al., 2010).

Experimental/Sleep Deprivation Phase. Upon completion of the Tuesday morning Practice Phase, subjects were released with instructions to return at 0530 Wednesday morning. Subjects were instructed to sleep according to their normal schedules, and to awaken at 0300 Wednesday, remaining awake until the 0530 report time. Compliance was gauged by actigraphy. Subjects were also re-familiarized with the protocol for the sleep deprivation phase of the study. Beginning at 0600 subjects were assessed on the PVT and the FSPT once every three (3) hours as follows. Trials began at 0600, 0900, 1200, 1500, 1800, 2100, 0000 (Thursday), and 0300. Upon completion of the final trial, subjects were debriefed and provided transportation to lodging facilities with instructions to obtain adequate sleep prior to check out.

Data Capture Set-up. The Fitness-to-Fly Assessment data collection software system uses a centralized server to consolidate data and enable easy extraction of study results from one source. The main component in the system from an operational viewpoint is the Data Capture Server (DCServer). The DCServer provides an easy-to-use graphical user interface (GUI) to facilitate data capture from the flight simulator workstations as well as facilities to review, modify and export the data to an Excel workbook for further analysis. It uses Microsoft Access as a data repository.

The flight simulator workstations are configured to run the X-Plane flight simulator software with a custom "plug-in" that was developed to dynamically capture in-flight data points such as longitude, latitude, heading, airspeed and altitude in real time. The plug-in transmits the captured data points to the DCServer where they are inserted into the data repository. The data collection activities run as a separate background thread to minimize impact on FSPT performance.

The DCServer and flight simulator workstations are connected via a simple local area network and communications between the machines is accomplished using TCP/IP Sockets with a simple protocol designed specifically for this application. The protocol provides for automatic workstation identification to the DCServer and allows the DCServer to control the starting and stopping of data capture for each workstation individually.

All software was developed with Microsoft Visual Studio 2008 using a combination of the C, C++ (for the plug-in) and C# (for the DCServer) languages.

FSPT Performance Metric Extraction. A series of Visual Basic (VB) programs were written as a macro workbook in Microsoft Excel to calculate basic descriptive statistics and lapse times for subsequent analyses. The macro workbook was comprised of eight worksheets from Test 1 (T1) to Test 8 (T8) of the Experimental/Sleep Deprivation Phase, as well as a central “control panel” sheet (with buttons linked to the individual macros) and an output worksheet for results. In addition, a record macro option was used to reformat the data sheets and to transform data as necessary for use in other statistical packages (i.e., SPSS and HLM). Using the workbook, descriptive statistics (mean, standard deviation, and variance) were calculated for altitude, heading, and airspeed. The FSPT outputs altitude in meters; these values were converted to feet for ease of translation with standard flight nomenclature. It was also necessary to transform the FSPT heading output. For example, in the standard compass configuration used, due North is equal to 0º degrees, and the heading increases in a clockwise direction, going through 180º at due South and arriving back at 360º or 0º at due North. Subjects in this experiment were instructed to fly due North. Thus, heading data consisted of values that were either slightly above 0 or slightly below 360. Given this, we transformed the heading data using a conditional statement: if the value is greater than 180, subtract 360; if the value is less than 180, do nothing. No transformation was necessary for airspeed, outputted in knots (kts). The Excel macro code for flight simulator data reduction can be found in the Appendix.

ANALYSES AND RESULTS

Overview

Initial psychometric properties of the FSPT were evaluated using test-retest reliability, construct validity, and convergent validity. Test-retest reliability was quantified using a combination of inter-trial correlations, inter-class correlations (ICC), and Cronbach’s Alpha of subject performance in order to establish the tool’s stability over time. Construct validity was established by testing performance across time employing a Repeated Measures Analysis of Variance (ANOVA), with the assumption that significant change in performance across testing sessions indicates conceptual sensitivity to fatigue. Convergent validity was established with Hierarchical Linear Modeling (HLM), using FSPT scores across time to predict concurrent PVT lapses across time.

FSPT Dependent Variable Preparation. Deviations from the specified flight parameter goals for heading (due North), airspeed (140 kts), and elevation (2000 ft) were calculated separately. Lapse times were calculated for each parameter as the number of seconds during a simulator trial that subjects deviated from the flight goal by greater than one intraindividual standard deviation (determined at baseline). Flight Simulator Total Lapse Time (FS Lapse) was the sum of lapse times for each parameter.

Test-Retest Reliability

A correlation matrix among FSPT lapse times at each Experimental/Sleep Deprivation Phase test time is shown in Table 2. Results indicate significant inter-trial correlations for FSPT performance were moderately strong (r = .53 - .81, $p < 0.05$), providing initial evidence for test-retest reliability; however, there were some notable non-significant correlations. For example, T1 displayed a significant relation with T2 and T5 only (r = .556, $p < 0.05$ and r = .558, $p < 0.05$, respectively). There are several factors that may explain this variability, many of which are key to analyzing the data in greater depth. First, and most importantly, a similar pattern emerges from our sample with PVT lapses, with the majority of inter-trial relations significantly correlated (see Table 3). However, as with FS Lapse, there are notable exceptions to this rule in which inter-trial relations are weak and non-significant. This initially surprising pattern for a gold-standard task is due to unusually high, yet stable, inter-individual variability in the measure. In the case of the PVT, consistent, trait-like inter-individual variability has been previously documented (Van Dongen, Maislin & Dinges, 2004; Chandler et al., 2010). As a result, traditional Pearson product moment correlation only captures part of the measure's true relation with itself across time and between individuals, treating all error variance as the same. Van Dongen, Maislin and Dinges (2004) explain that when measuring reactions to a dynamic stressor, such as fatigue, within and between subject error variance must be considered separately and then compared. The authors suggest using ICC to identify the amount of variance that may be attributed to stable inter-individual variability. If that amount is significant, then theoretically unclear gaps in a traditional Pearson product moment correlation matrix can be satisfactorily explained. For example, Van Dongen and colleagues cite an ICC value of 0.58 (F [9, 9] = 4.61, p = .016) for PVT lapses in a sample of healthy adults, meaning that 58% of the variance in PVT lapses can be attributed to stable inter-individual differences in fatigue susceptibility (2004). In the current sample, FS Lapse time has an ICC value of 0.48 (F [14, 14] = 8.33, $p < 0.01$), indicating that 48% of the variance in the measure is due to between-subjects variability, thus explaining the attenuated correlations when using Pearson's r. As a final step, the overall reliability of the measure may be clarified using Cronbach's Alpha, considered in combination with the ICC. In the current sample, FS Lapse time has a Cronbach's Alpha of .89, a high value for an applied measure (Nunnally, 1978).

With thorough analysis, the FSPT displayed strong initial reliability. The exact nature of that reliability, with a significant amount of stable between-subjects variance included, demonstrates the need for mixed-modeling analysis in characterizing aspects of the task's validity as well, considered next.

Table 2. Inter-trial Correlations of FSPT Total Lapse Times Using Pearson Product-Moment Correlation

		1	2	3	4	5	6	7	8
1	Test 1	–							
2	Test 2	0.556 *	–						
3	Test 3	0.438 –	0.732 **	–					
4	Test 4	0.084 –	0.337 –	0.566 *	–				
5	Test 5	0.558 *	0.741 **	0.807 **	0.46 –	–			
6	Test 6	0.342 –	0.528 *	0.694 **	0.698 **	0.477 –	–		
7	Test 7	0.257 –	0.329 –	0.564 *	0.635 *	0.329 –	0.847 **	–	
8	Test 8	0.263 –	0.414 –	0.4 –	0.648 **	0.259 –	0.785 **	0.781 **	–

Note: ** $p < .01$, * $p < .05$

Table 3. Inter-trial Correlations of PVT Lapses Using Pearson Product-Moment Correlation

		1	2	3	4	5	6	7	8
1	Test 1	–							
2	Test 2	0.772 **	–						
3	Test 3	0.71 **	0.651 **	–					
4	Test 4	0.595 *	0.574 *	0.944 **	–				
5	Test 5	0.686 **	0.731 **	0.577 *	0.533 *	–			
6	Test 6	0.215 –	0 –	0.166 –	0.495 –	0.205 –	–		
7	Test 7	0.304 –	0.346 –	0.809 **	0.819 **	0.435 –	0.476 –	–	
8	Test 8	0.509 –	0.442 –	0.937 **	0.945 **	0.498 –	0.437 –	0.914 **	–

** $p < .01$, * $p < .05$

Construct Validity

PVT Lapses. PVT lapses per trial were analyzed across time using repeated measures ANOVA. Results indicate significant fatigue effects for lapses, replicating previously established patterns (see Table 4 and Figure 1). Results are presented here for ease of comparison to FS Lapse.

Table 4. ANOVA results for PVT

	F	df	*p*	η_p^2
PVT Lapses[†]	6.88	(1.45, 20.28)	< 0.01	.329

[†] Geisser-Greenhouse correction used due to violation of sphericity

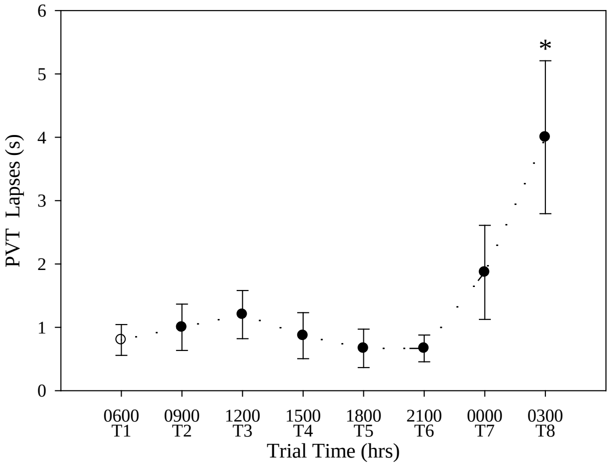


Figure 1. Mean PVT Lapses at each test trial across time. Post-hoc analyses revealed significant differences between T8 and all other trials, indicating a distinct point at which group vigilance began to fail (*).

FSPT Total Lapse Time. The analysis revealed significant effects of assessment time on total lapse time, suggesting that total lapse time is sensitive to fatigue effects. Results are displayed in Table 5 and Figure 2. Notably, this pattern is similar to PVT lapses (see Figure 1).

Table 5. ANOVA results for FSPT Total Lapse Time[†]

F	df	*p*	η_p^2
4.45	(2.64, 36.90)	< 0.05	.241

[†] Geisser-Greenhouse correction used due to violation of sphericity

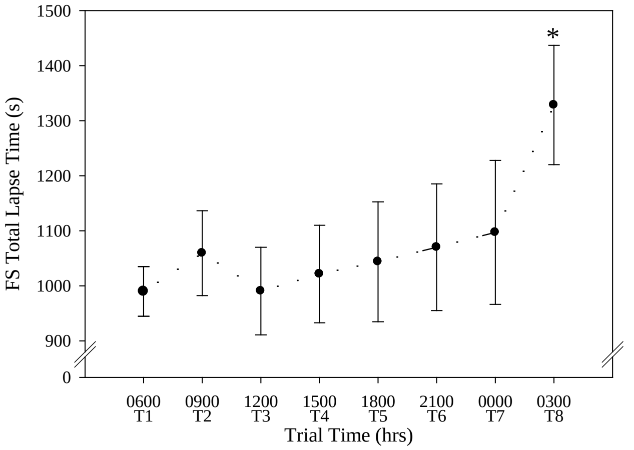


Figure 2. Mean Total Lapse Time on FSPT scores in seconds at each test trial across time. Post-hoc analyses revealed significant differences between T8 and all other trials, indicating a distinct point at which group vigilance began to fail (*). Notably, this pattern is similar to PVT lapses (Figure 1).

Convergent Validity

Based on the significant between-subjects variability identified in the reliability analyses, a bivariate Hierarchical Linear Model was used to examine the ability of FS Lapses to predict concurrent PVT lapses as a test of convergent validity. The use of HLM allows simultaneous examination of group and individual relations between the two lapse counts. Both fixed (level 1 equations) and random (level 2 equations) effects of the predictor were included, allowing determination of an overall effect of FS Lapses on PVT Lapses, as well as whether the relation was consistent or varied across subjects. Significance at level 1 indicates a group effect (within-subjects variability); while significance at level 2 indicates significant individual differences within that overall effect (between-subjects variability).

Level 1 and level 2 equations were significant for FS Lapses predicting PVT Lapses. The significant level 1 relation indicates that as FS Lapse time increases, PVT lapses increase. Visual inspection of the significant inter-slope variability at level 2 shows that for some individuals, this relation is nearly 1:1, with lapses on both measures progressing at similar rates (or not progressing at all), while for others the breakdown in performance is more pronounced for one task compared to the other (see Figure 3). For example, long, flat lines at the bottom of Figure 3 represent individuals for whom there was significant variability in FS Lapse with little variability in PVT lapses. The significant level 2 effect confirms the presence of stable inter-individual variability while the level 1 effect provides evidence of convergent validity.

Table 6. Bivariate HLM Relation: FSPT Total Lapse Time Predicting Outcome PVT Lapses

	Level 1				Level 2			
Variable	Equation	t	df	*p*	Equation	χ2	df	*p*
FS_Total	Y = P0 + P1*(FS_LAPSE) + E	3.386	14	<0.01	P0 = B00 + R0 P1 = B10 + R1	142.49645	14	<0.01

Note: PVT = Psychomotor Vigilance Task, FS_Total = FSPT Total Lapse Time

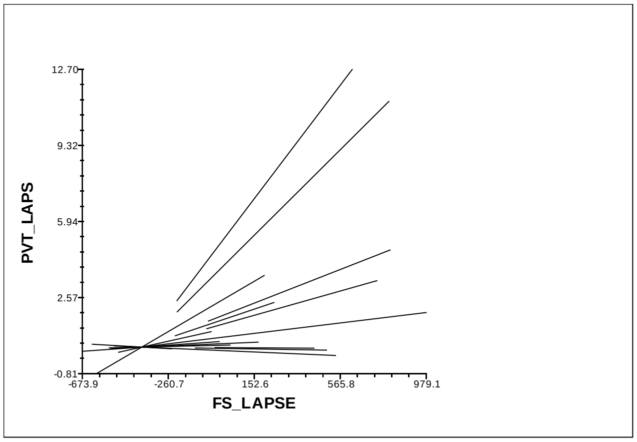


Figure 3. Individual slopes for PVT Lapses in relation to FSPT Total Lapse Time (group mean centered values). There was a significant group effect and significant individual differences such that, on average, an increase in FSPT Total Lapse Time translated into an increase in PVT Lapses; however, the nature of that relation varied significantly from subject to subject.

DISCUSSION

The PVT is the most widely accepted standard for measuring cognitive performance in fatigue studies. This paper presents preliminary evidence that an ecologically valid flight simulation task can be used with similar diagnosticity, validity, and reliability as the PVT for assessing fatigue in Naval Aviators. The use of a simple, off-the-shelf flight simulation task to measure flight simulator performance while fatigued represents an attempt to bridge the ever-present gap between experimental control and ecological validity in the laboratory. This gap, known as the “artificiality criticism” in experimental psychology, has long been debated with strong arguments for both approaches (e.g., Henshel, 1980). Rather than polarizing our efforts within the artificiality argument, we sought to strike a balance between psychometric strength and ecological practicality. The PVT is an indispensible research tool; however, we argue that increased realism in applied research, especially when dealing with a specific situation such as fatigue in flight, can only benefit the advancement of operational knowledge. The FSPT was purposefully constructed to measure performance under fatigue similarly to the PVT while capturing physical aspects of the environment to which its result will be applied: flight performance. Dorrian, Rogers, and Dinges provide eight criteria for any neurocognitive assay purporting to assess the effects of sleep deprivation (2005, p.42, Table 1). The current evidence suggests that the FSPT meets, or has the potential to meet, these criteria, as follows.

The task must reflect a fundamental aspect of waking neurocognitive functions. We designed the FSPT as a monitoring and reaction test in order to tap basic attentional processes affected by fatigue. Participants were required to monitor airspeed, heading, and altitude, and make corrections to each of these in order to maintain specified parameters. The basic structure of the task is purposefully analogous to the PVT, where individuals must monitor the screen for changes (i.e., the appearance of a number) and react by making a correction (pushing the response button). In both cases, the goal state is a return to specified parameters as quickly as possible once the parameters are exceeded. This type of performance is a fundamental aspect of successful neurocognitive functioning when awake, with the FSPT capturing it in a specific setting.

The task must be suitable for repeated administrations. Dorrian and colleagues further explain this criterion by specifying that the task must have “…a minimal learning curve” (2005, p.42). The design of the larger study included 2 days of baseline testing in order for participants to practice all tasks to asymptote. At the beginning of each FSPT session, participants were allowed unlimited time to reach the flight parameters. Once reached, participants would give a “thumbs-up”, and data collection would begin. According to study notes, time to data collection start decreased dramatically across the two baseline days, and then remained steady during continual wakefulness. Asymptote was then confirmed via visual inspection of individual performance slopes across the entire study. Although, unlike the PVT, initial practice is required, time to FSPT asymptote is sufficiently short in the current sample to accommodate a brief fatigue protocol (2 hours of baseline testing). Significant decrement across time awake, without performance increase above asymptote, confirms the task’s suitability for a repeated measures design.

The task must be easily performed with no aptitude effects. Specifically, there must be evidence that the task “[1] yields consistent results among a wide range of subject populations, [2] can be taught quickly, and [3] can be used in laboratory experiments, simulator scenarios, and field situations” (Dorrian et al., 2005, p.42). The current report discusses the first, and to our knowledge only, administration of this exact form of the FSPT. Therefore, points 1 and 3 cannot be directly addressed yet, and will be the subject of follow-on work. Point 2 is supported by the quickness with which participants reached performance asymptote (see above). Despite the fact that our subjects were training to be Naval Aviators, only 2 had actual prior flight experience, meaning that the task was easily taught to novice flyers.

The task duration must be relatively brief. The FSPT is 15 minutes long, which compares well to the most widely used duration of the PVT (10 minutes). Work on shorter PVT durations (5 and 2 minutes) highlights the importance of using at least a 10-minute time frame for lapse analysis (Loh, Lamond, Dorrian, Roach & Dawson, 2004), since sustained time-on-task is a fundamental component of capturing performance decrement due to fatigue. The FSPT was easily integrated into a repeated-measures design which included several other tests, including the PVT itself, meeting the criterion that it “not result in greatly augmented subject burden” (Dorrian et al., 2005, p.42). Even so, future work will focus on testing multiple durations of the FSPT, beginning with the 10-minute PVT time frame.

The task must have a high signal load for analysis. The requirement to “provide[s] a large number of behavioral samples in a brief period of time” is a strength of the FSPT. It has an extremely high signal load for its length, sampling deviation from each goal parameter every second for the duration of the task, equaling 900 samples per parameter, or 2,700 samples total. Each sample represents the opportunity for a lapse, quantified as deviation from the flight goal by greater than one intraindividual standard deviation within the sampled second. High signal load gives the FSPT extreme sensitivity to the effects of fatigue-induced wake state instability and momentary performance compensation, one of the neurocognitive hallmarks of sleep deprivation (see Durmer and Dinges, 2005).

The task must be reliable. Dorrian and colleagues state that the task “[1] challenges the subject to maintain cognitive output, [2] provides test-re-test stability, and [3] reflects trait-like inter-individual differences” (2005, p. 42). The act of constantly monitoring and actively adjusting 3 flight parameters, and the fatigue related decrements across time we have documented on those actions, satisfies point 1. Point 2 was also addressed in the analyses, with inter-trial test-retest reliability falling in the moderate to high range (r = .53 - .81, all p < .05). Point 3 is addressed by the presence of a significant ICC value for FS Lapse (.48), and significant level 2 effects using HLM (see Table 6), indicating that there are stable inter-individual differences on the task in response to fatigue.

The task must be valid. This criterion includes subsets of convergent, ecological, and theoretical validity. The type of convergent validity discussed by Dorrian and colleagues, sensitivity “to many forms of sleep deprivation” (2005; p. 42) cannot yet be established, though the task is currently being tested in a chronic sleep restriction in our lab for that purpose. However, HLM analyses did reveal the FSPT’s ability to predict simultaneous PVT performance, offering convergent validity through covariation with an established tool. Demonstrating ecological validity specific to flight performance was the main goal of constructing the FSPT, one that was achieved simply by using a flight simulator, rather than a basic cognitive task, as a measure of performance while fatigued. Theoretical, or construct, validity was specifically addressed in the GLM analysis. Significant performance decrements on the FSPT were observed across time of continual wakefulness ($p < 0.05$).

Results can be interpreted in a meaningful way. Meeting this criterion represents the strongest aspect of the FSPT. Total lapse time in seconds is literally the amount of time that flight parameters are not being met successfully. Decisions in the cockpit are often made in milliseconds. Operationally, this translates into the likelihood of a mishap occurring; the longer the total lapse time, the larger the window for critical mistakes.

Next Steps / Future Directions

The FSPT is currently being used in a chronic sleep restriction study with a population of Naval aircrew students. Building on the current results, we will continue to employ the task in different fatigue situations with different military aviation populations to fully test its psychometric and theoretical profile.

Summary

Preliminary evidence suggests that performance on a simple flight simulation task can be used as a reliable, ecologically valid measure of fatigue in Student Naval Aviators. Future work should focus on replication and extension of the FSPT to further establish the task's promising psychometric properties and operational applications.

Acknowledgements: The authors would like to thank CDR Rita Simmons, Ph.D., N. Summer Dodson, M.S., Ashley Turnmire, M.S., Andrew Reinert, M.S. and Jill Bennett, M.S., for thoughtful contributions to drafts of this manuscript.

Disclaimers: The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. I am an employee of the U.S. Government. This work was prepared as part of my official duties. Title 17, USC, §105 provides that 'Copyright protection under this title is not available for any work of the U.S.Government.' Title 17, USC, §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. The study protocol was approved by the Naval Aerospace Medical Research Laboratory Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects. Approved for public release; distribution is unlimited. This work was supported by funded by work unit number 70803.

References

Adamson, M. M., Samarina, V., Xiangyan, X., Huynh, V., Kennedy, Q., Weiner, M., …Taylor, J. (2010). The impact of brain size on pilot performance varies with aviation training and years of education. *Journal of International Neuropsychological Society*, *16*, 412–423. doi:10.1017/S1355617710000111

Balkin, T. J., Bliese, P. D., Belenky, G., Sing, H., Thorne, D. R., Thomas, M., …Wesensten, N. J. (2004). Comparative utility of instruments for monitoring sleepiness-related performance decrements in the operational environment. *Journal of Sleep Research*, *13*, 219-227.

Caldwell, J. A., Caldwell, J. L., Brown, D. L., & Smith, J. K. (2004). Modafinil's effects on simulator performance and mood in pilots during 37 h without sleep. *Aviation, Space, and Environmental Medicine, 75*, 777-784.

Caldwell, J. A., Caldwell, J. L., Brown, D. L., Smythe, N. K., Smith, J. K., Mylar, J., …Schroeder, C. (2003). The effects of 37 hours of continuous wakefulness on the physiological arousal, cognitive performance, self-reported mood, and simulator flight performance of F-117A pilots. *Technical Report* AFRL-HE-BR-TR-2003-0086. 75 pages.

Caldwell, J. A., Mallis, M. M., Caldwell, J. L., Paul, M. A., Miller, J. C., & Neri, D. F. (2009). Fatigue countermeasures in aviation. *Aviation, Space, and Environmental Medicine, 80*, 29-59.

Caldwell, J. A., Smythe, N. K., Leduc, P. A., & Caldwell, J. L. (2000). Efficacy of Dexedrine for maintaining aviator performance during 64 hours of sustained wakefulness: a simulator study. *Aviation, Space, and Environmental Medicine, 71*, 7-18.

Chandler, J. F., Arnold, R. D., Phillips, J. B., Lojewski, R. A., & Horning, D. S. (2010). *Preliminary validation of a readiness-to-fly assessment tool for use in Naval aviation* (Report No. ADA522106). Retrieved from http://www.handle.dtic.mil/100.2/ADA522106

Dalecki, M., Bock, O., & Guardiera, S. (2010). Simulated flight path control of fighter pilots and novice subjects at +3G_z in a human centrifuge. *Aviation, Space, and Environmental Medicine, 81*, 484-488.

Dinges, D. F., & Powell, J. W. (1985)**.** Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations**.** *Behavior Research Methods, Instruments, & Computers 17*, 652-655**.**

Dorrian, J., Rogers, N. L., & Dinges, D. F. (2005). Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In Kushida, C.A. (Ed.), *Sleep Deprivation: Clinical Issues, Pharmacology and Sleep Loss Effects* (pp. 39-70). New York, NY: Marcel Dekker, Inc.

Durmer, J. S., & Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. *Seminars in Neurology, 25*, 117-129.

Henshel, R. L. (1980). The purpose of laboratory experimentation and the virtues of deliberate artificiality. *Journal of Experimental Social Psychology, 16*, 466-4778.

Killgore, W. D., Grugle, N. L., Reichardt, R. M., Killgore, D. B., & Balkin, T. J. (2009). Executive functions and the ability to sustain vigilance during sleep loss. *Aviation, Space, and Environmental Medicine, 80),* 81-87.

Leino, T. K., Lohi, J. J., Huttunen, K. H., Lahtinen, T. M., Kilpeläinen, A. A., & Muhli, A. A. (2007). Effect of caffeine on simulator flight performance. *Military Medicine, 172,* 982-987.

Loh, S., Lamond, N., Dorrian, J., Roach, G., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, 36,* 339-346.

Naval Safety Center (2006). Aeromedical factors involved in Naval aviation class A flight mishaps from FY00-06. Unpublished briefing slide. Norfolk, VA.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Russo, M. B., Kendall, A. P., Johnson, D. E., Sing, H. C., Thorne, D. R., Escolas, S. M., …Redmond, D. P. (2005). Visual perception, psychomotor performance, and complex motor performance during an overnight air refueling simulated flight. *Aviation, Space, and Environmental Medicine, 76*(Suppl. 7), C92-C103.

Van Dongen, H. P., Baynard, M. D., Maislin, G., & Dinges, D. F. (2004). Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. *Sleep, 27,* 423-33.

Van Dongen, H. P., Caldwell, J. A., & Caldwell, J. L. (2006). Investigating systematic individual differences in sleep-deprived performance on a high-fidelity flight simulator. *Behavior Research Methods, 38,* 333-343. DOI: 10.3758/BF03192785

Van Dongen, H. P., Maislin, G., & Dinges, D. F. (2004). Dealing with inter-individual differences in the temporal dynamics of fatigue and performance: Importance and techniques. *Aviation, Space, and Environmental Medicine, 75*(Suppl. 3), A147-A154.

APPENDIX. Excel Macro Code for Flight Simulator Data Reduction

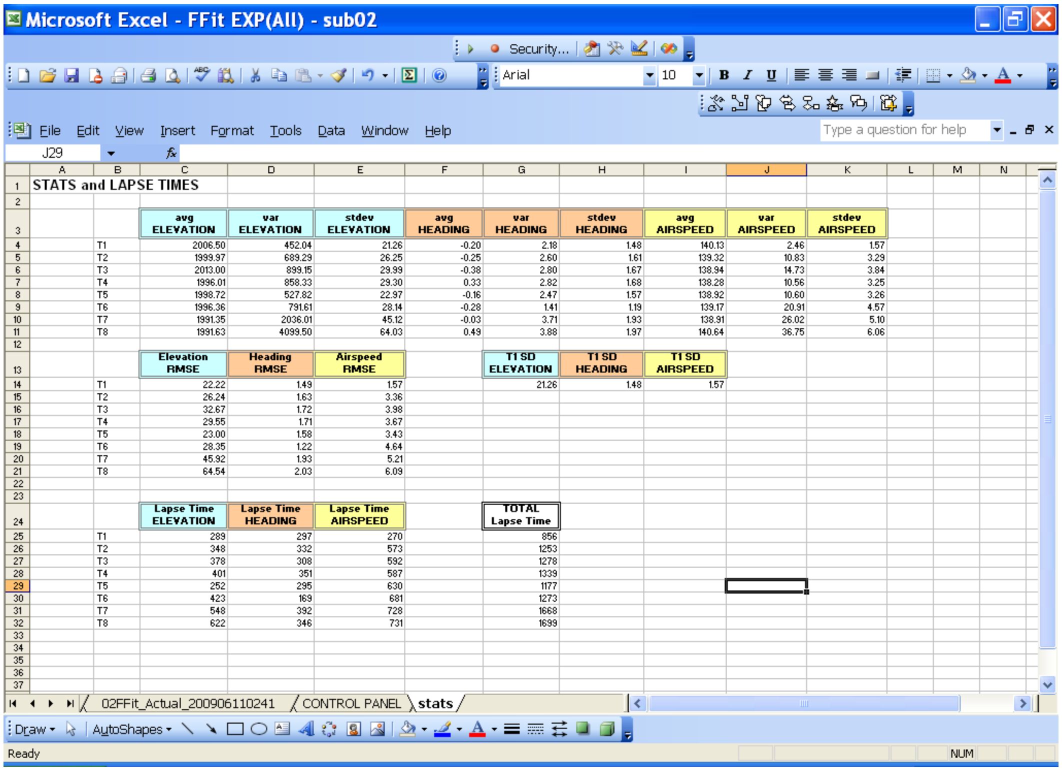
Visual Basic (VB) Code:

- Range("K2").Select
- ActiveCell.FormulaR1C1 = "=IF(RC[-3]>180,RC[-3]-360,RC[-3])"
- Range("K2").Select
- Selection.AutoFill Destination:=Range("K2:K901")

Figure A1. Code for calculating statistics on performance in maintaining elevation

```
*****************************************************************************************
Sub Calc_Stats()

Dim i, s As Integer
Dim dRange As Range

s = 1
For s = 1 To 8

i = 3
Sheets(s).Select

Set dRange = Sheets(s).Range("k2:k901")

   Average = WorksheetFunction.Average(dRange)
   Sheets("stats").Cells(s + 3, i).Value = Average

   Variance = WorksheetFunction.Var(dRange)
   Sheets("stats").Cells(s + 3, i + 1).Value = Variance

   StandardDev = WorksheetFunction.StDev(dRange)
   Sheets("stats").Cells(s + 3, i + 2).Value = StandardDev

Next s

End Sub
*****************************************************************************************
```

Figure A2.


Microsoft Excel - FFit EXP(All) - sub02
File Edit View Insert Format Tools Data Window Help
J29
Arial
10
Type a question for help
02FFit_Actual_200906110241
CONTROL PANEL
stats
Draw
AutoShapes
Ready
NUM

STATS and LAPSE TIMES

	avg ELEVATION	var ELEVATION	stdev ELEVATION	avg HEADING	var HEADING	stdev HEADING	avg AIRSPEED	var AIRSPEED	stdev AIRSPEED
T1	2006.50	452.04	21.26	-0.20	2.18	1.48	140.13	2.46	1.57
T2	1999.97	689.29	26.25	-0.25	2.60	1.61	139.32	10.83	3.29
T3	2013.00	899.15	29.99	-0.38	2.80	1.67	138.94	14.73	3.84
T4	1996.01	858.33	29.30	0.33	2.82	1.68	138.28	10.56	3.25
T5	1998.72	527.82	22.97	-0.16	2.47	1.57	138.92	10.60	3.26
T6	1996.36	791.61	28.14	-0.28	1.41	1.19	139.17	20.91	4.57
T7	1991.35	2036.01	45.12	-0.03	3.71	1.93	138.91	26.02	5.10
T8	1991.63	4099.50	64.03	0.49	3.88	1.97	140.64	36.75	6.06

	Elevation RMSE	Heading RMSE	Airspeed RMSE		T1 SD ELEVATION	T1 SD HEADING	T1 SD AIRSPEED
T1	22.22	1.49	1.57		21.26	1.48	1.57
T2	26.24	1.63	3.36				
T3	32.67	1.72	3.98				
T4	29.55	1.71	3.67				
T5	23.00	1.58	3.43				
T6	28.35	1.22	4.64				
T7	45.92	1.93	5.21				
T8	64.54	2.03	6.09				

	Lapse Time ELEVATION	Lapse Time HEADING	Lapse Time AIRSPEED		TOTAL Lapse Time
T1	289	297	270		856
T2	348	332	573		1253
T3	378	308	592		1278
T4	401	351	587		1339
T5	252	295	630		1177
T6	423	169	681		1273
T7	548	392	728		1668
T8	622	346	731		1699

Figure A3. Code for measuring lapses in maintaining elevation

```
*****************************************************************************************
Sub Lapse_Calculator()

Dim s As Integer
Dim i As Integer

ThisWorkbook.Sheets("stats").Activate
bound = [T1_SD_Elevation].Value

s = 1
For s = 1 To 8
Sheets(s).Select

lapse = 0
i = 2
For i = 2 To 901
   If Sheets(s).Cells(i, "K").Value > 2000 + bound Then
      lapse = lapse + 1
```

```
    End If
    If Sheets(s).Cells(i, "K").Value < 2000 - bound Then
        lapse = lapse + 1
    End If
Next i

    Sheets("stats").Cells(s + 24, 3).Value = lapse

Next s

End Sub
**********************************************************************************************
```

REPORT DOCUMENTATION PAGE

1. REPORT DATE (DD MM YY)
X 12 August 2011

2. REPORT TYPE
X Technical Report

3. DATES COVERED (from – to)
X 1 Jan 2009 – 31 Dec 2011

4. TITLE
X The use of commercial flight simulation software as a psychometrically sound, ecologically valid measure of fatigued performance

5a. Contract Number:
5b. Grant Number:
5c. Program Element Number:
5d. Project Number:
5e. Task Number:
5f. Work Unit Number:70803

6. AUTHORS
X Chandler, J. F., Horning, D. S., Arnold, R. D., Phillips, J. B., Horak, D. S., Taylor, D. L.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Medical Research Unit – Dayton
2624 Q St., Bldg. 851, Area B
WPAFB, OH 45433-7955

8. PERFORMING ORGANIZATION REPORT NUMBER
NAMRU-D-11-34

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Bureau of Medicine and Surgery Medical Development Program
Department of the Navy
2300 E Street, NW
Washington, DC 20372-5300

10. SPONSOR/MONITOR'S ACRONYM(S)
BUMED

11. SPONSOR/MONITOR'S REPORT NUMBER(s)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
X **A brief (200 words max) factual *summary* of the most significant information.**

Fatigue is a deadly problem for U.S. Naval Aviation (Naval Safety Center, 2006), and receives a correspondingly large amount of research attention in military RDT&E. Though the basic consequences of fatigue are well known, significant measurement challenges remain in the applied laboratory, where an optimal combination of scientific rigor and operational relevance can be elusive. The purpose of this report is twofold: 1) to describe the development and execution of a flight simulation tool for quantifying vigilance during a fatigue study, and 2) to describe the effort to balance the control and diagnosticity of the PVT with the ecological validity of flight simulation in an inexpensive, off-the-shelf, open-source format. Fifteen active duty military personnel from the Naval Aviation Preflight Indoctrination (API) program at NAS Pensacola volunteered for the study. Subjects completed a battery of neurocognitive and physiological assessments over the course of 25 hours of continual wakefulness. Significant inter-trial correlations for the Flight Simulator Performance Task (FSPT) scores were moderately strong (r = .53 - .81, p < .05), providing initial evidence for test-retest reliability. However, there were some notable non-significant correlations which are discussed in terms of significant, trait-like individual differences in the data. Preliminary evidence suggests that performance on a simple flight simulation task can be used as a reliable, ecologically valid measure of fatigue in Student Naval Aviators. Future work should focus on replication and extension of the FSPT to further establish the measure's psychometric properties.

15. SUBJECT TERMS
X fatigue, flight simulation, PVT, X-plane, vigilance, continual wakefulness, psychometrics

16. SECURITY CLASSIFICATION OF:
a. REPORT UNCL
b. ABSTRACT UNCL
c. THIS PAGE UNCL

17. LIMITATION OF ABSTRACT
UNCL

18. NUMBER OF PAGES
20

18a. NAME OF RESPONSIBLE PERSON
Commanding Officer

18b. TELEPHONE NUMBER (INCLUDING AREA CODE)
COMM/DSN: (937) 938- 3872